What Synthesis Is Missing: Depth Adaptation Integrated With Weak Supervision for Indoor Scene Parsing

Keng-Chi Liu, Yi-Ting Shen, Jan P. Klopp, Liang-Gee Chen
National Taiwan University
{calvin89029,dennis45677,kloppjp}@gmail.com, lgchen@ntu.edu.tw

Abstract

Scene Parsing is a crucial step to enable autonomous systems to understand and interact with their surroundings. Supervised deep learning methods have made great progress in solving scene parsing problems, however, come at the cost of laborious manual pixel-level annotation. Synthetic data as well as weak supervision have been investigated to alleviate this effort. Nonetheless, synthetically generated data still suffers from severe domain shift while weak labels often lack precision. Moreover, most existing works for weakly supervised scene parsing are limited to salient foreground objects. The aim of this work is hence twofold: Exploit synthetic data where feasible and integrate weak supervision where necessary. More concretely, we address this goal by utilizing depth as transfer domain because its synthetic-to-real discrepancy is much lower than for color. At the same time, we perform weak localization from easily obtainable image level labels and integrate both using a novel contour-based scheme. Our approach is implemented as a teacher-student learning framework to solve the transfer learning problem by generating a pseudo ground truth. Using only depth-based adaptation, this approach already outperforms previous transfer learning approaches on the popular indoor scene parsing SUN RGB-D dataset. Our proposed two-stage integration more than halves the gap towards fully supervised methods when compared to previous state-of-the-art in transfer learning.

1. Introduction

Scene parsing is an important computer vision task aiming at assigning semantic information to the entire image and providing a complete understanding of the scene. State-of-the-art scene parsing works [7, 23, 24, 46] heavily rely on human labeled pixel-level data, which is expensive and cumbersome to collect. To enable computer vision applications without such labeling efforts, two paradigms have been investigated: unsupervised domain adaptation and weak supervision. Domain adaptation for scene parsing (c.f. [15]) addresses the problem by transferring from a source domain (simulation) to features that are aligned with target domain (real data) without any labeled target samples. In spite of the progress that has been accomplished in realistic scene rendering and transfer learning approaches, there is still a significant domain discrepancy between real and synthetic imagery, especially in texturing. Weak supervision tackles this issue by leveraging weak annotations with lower acquisition costs such as bounding boxes [10, 21, 26, 27], scribble [22], points [3] or even image-level labels [1,6,19,20,27,30–33,35,38,41–43]. This enables a more cost-effective scaling of training datasets. Nevertheless, for image-level annotations, issues such as lack of boundary information, rare pixels for objects of interest, class co-occurrence and discriminative localization remain tremendous challenges. Moreover, the majority of existing works for weak supervision are only capable of handling salient foreground objects.

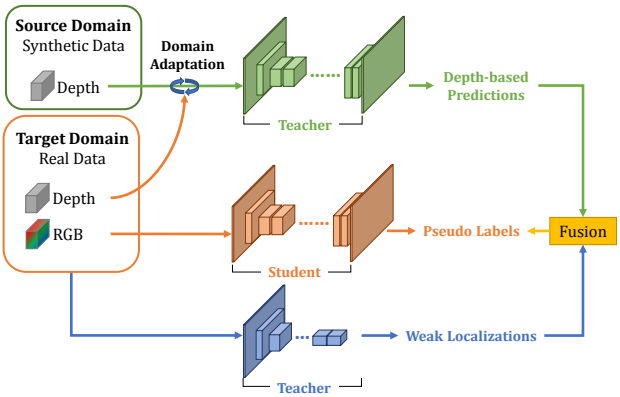

Figure 1. Illustration of our teacher-student framework. The teacher utilize depth as a low domain-shift auxiliary cue. This is fused with weak localization information to generate pseudo labels, which are used to train the student.

In this work, we aim at improving performance by transferring through a path of little domain discrepancy. While RGB images contain rich information, it is difficult to trans-

fer from synthetic to real instances in the RGB domain. Hence, we resort to depth information as an auxiliary cue that can be easily captured and is only used at training time. In the depth domain only the object geometry is of interest, which is easier to accurately synthesize and hence presents less domain shift. Therefore we adapt the depth cue to model sensing artifacts that are typically encountered in real depth measurements. However, the resulting teacher network is unable to segment all categories properly. Books in a book shelf, for example, do not have a distinctive geometry. To recover such information, we leverage image-level object tags. Such tags are easy to acquire, but do not come with location or boundary information. We hence adapt a weak localization technique to obtain heat maps from RGB images through a network trained solely on these image-level tags. Finally, the localization heat map information is fused with the depth-based predictions to yield a pseudo ground truth, which is in turn used to train the final student network on RGB images only. Fig. 1 illustrates our approach. Our main contributions can be summarized as follows:

- We propose a teacher-student learning procedure to learn scene parsing through low domain shift auxiliary cues and weak domain-specific annotations. The student network is shown to surpass its teacher, leading to 58% reduction of the gap between state-of-the-art supervised and domain adaptation methods.
- We are the first to perform depth map adaptation through cycle consistent adversarial networks, utilizing a min-max normalization to ensure proper learning of real depth map noise. It is shown to perform favorably against state-of-the-art domain adaption results on SUN RGB-D [39].
- A two-stage voting mechanism is proposed to integrate cues from depth adaptation and weak localization based on contour maps.

In order to facilitate low complexity mobile inference, we furthermore apply complexity reduction techniques to your final model. Related results are presented in the supplementary material as those are not our own contributions.

2. Related Work

2.1. Domain Adaptation

Domain adaptation aims at transferring source data to features that are aligned with the target domain so as to generalize the ability of the learned model and improve the performance on the task in target domain without target labels [15]. Recently, with the progress made in computer graphics, adaptation between synthetic and real domain has become a popular path for various computer vision tasks. Several datasets such as SceneNet [25], Pbrs [45] have been proposed for scene parsing. Unfortunately, severe domain shift is still met by virtue of the difficulties generating photo-realistic imagery. Therefore, several adaptation methods [8, 9, 15, 16, 44] have been proposed to reduce the simulation-to-real gap by means of Generative Adversarial Networks (GAN). [16] applies techniques of global and category specific adaptation. The global statistics are aligned by using a domain adversarial training technique. [8] extends the approach by not only aligning global statistics but class-specific ones as well. [9] uses the target guided distillation strategy from [14] and spatially-aware adaptation during the training process. [44] applies domain adaptation in a curriculum learning [4] fashion, learning scene parsing from tasks that are less sensitive to the aforementioned domain discrepancy. Moreover, [15] combines a cycle consistency reconstruction loss as proposed by [48] with a generative approach to prevent the mapping functions from contradicting each other.

2.2. Weak Image-level Supervision

Weakly supervised approaches leverage weak annotations that come at lower costs than the original ones. Since such annotations are efficient to collect, one can build a large-scale dataset for diverse semantic categories with less effort and learn scene parsing in the wild. Early works mostly applied methods based on graphical models which infer labels for segments with probability relations between images and annotations. Additionally, class-agnostic cues and post-processing are often used to improve the results. Among those methods exploiting only weak annotations, learning only from images is the most economical but also challenging one. Paradigms such as multiple instance learning (MIL) [2] and self-training [17] are often applied. [27] adopt a self-training EM-like procedure, where the model is recursively updated from the results created by itself. [31] formulates the task as a MIL problem by applying a global max pooling after the CNN to enforce the predictions correspond to positive classes. Recently, techniques based on discriminative localization [37, 47], which probe into the contribution of each hidden neuron, are often employed. SEC [19] uses such discriminative localization to indicate a position within the area of a semantic class and expand it to neighboring pixels. However, neural networks tend to focus only on discriminative parts and not on the object as a whole. Hence, works have been focusing on transferring information to the non-discriminative part of objects. [42] obtains improvements by exploiting an adversarial erasing method. Class-agnostic cues are used to obtain shape or instance information in most works that achieve state-of-the-art results [6]. [41] uses both techniques to mine common object features from the initial localization, ex-

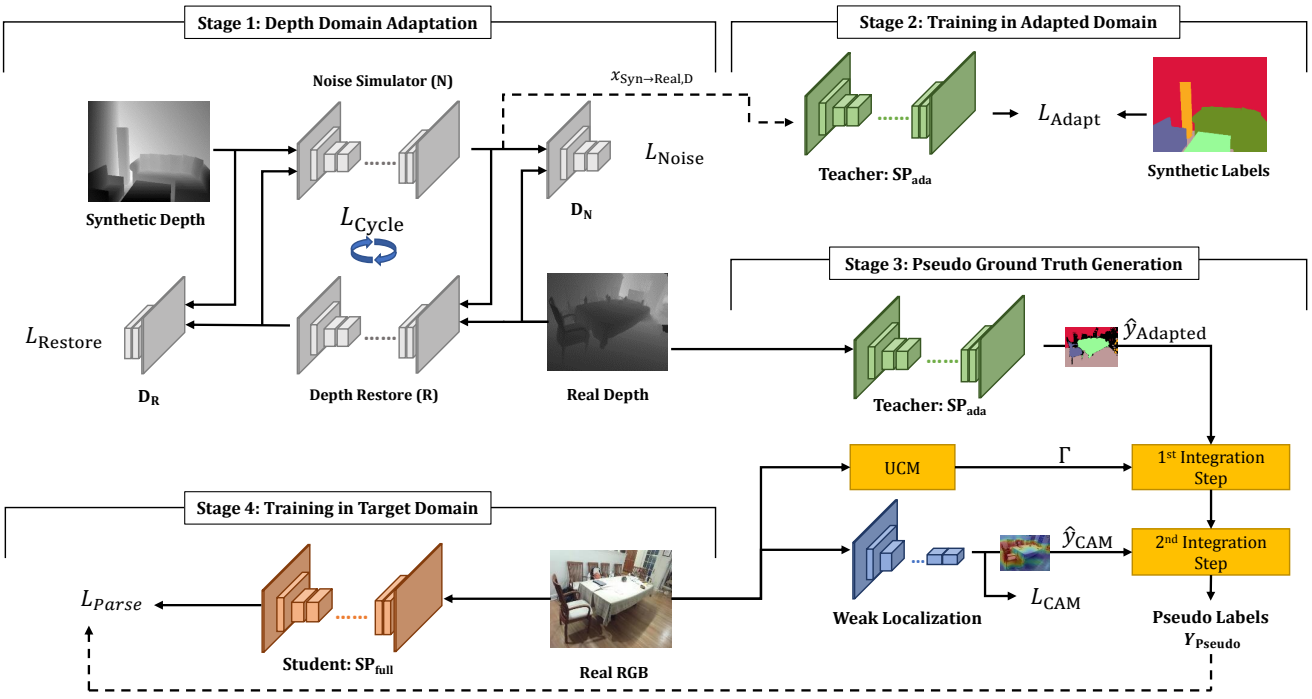

Figure 2. **Overview of our proposed framework.** A four stage design first adapts synthetic depth maps to appear like real ones. Those adapted depth maps are then used to train a teacher in stage two. Stage three fuses the teacher's predictions with weak localization from class activation maps (CAM) based on contour maps to generate pseudo ground truth. Finally, in stage four, the student network is trained on RGB data using the pseudo labels from the previous stage.

pand object regions and consider saliency maps under a Bayesian framework. [1] propagates semantic information by a random walk with the affinities predicted by AffinityNet. [43] argues that varying dilation rates can effectively promote object localization maps. Furthermore, most existing works are dedicated to handle multiple salient foreground instances and evaluate on the Pascal VOC dataset [12]. [36] is the only existing work that considers complete scene parsing (background + foreground) with only image-level label by leveraging two-stream deep architecture and heat map loss. However, their result does not perform well compared to other adaptation methods on the Cityscape dataset.

3. Proposed Method

In this section, we present the details of our proposed scene parsing framework. Fig. 2 illustrates how it proceeds in four stages: First, we adapt the depth cues from the synthetic into the real domain. Second, we train a teacher network on the adapted synthetic depth cues. Third, by applying the teacher network to the target (real) domain and integrating the generated labels with weak localization over contour maps, we obtain robust pseudo ground truth. Lastly, we train the student network on the target domain RGB input using the constructed pseudo ground truth.

3.1. Depth Domain Adaptation

Our objective is to transfer label information from synthetic data $X_{\text{Syn}} = \{X_{\text{Syn,D}}\}$ to the real domain $X_{\text{Real}} = \{X_{\text{Real,D}}, X_{\text{Real,RGB}}\}$ while only using depth cues due to their smaller domain gap when compared to RGB. While it is possible to train on synthetic depth data directly, the domain gap still leads to noticeable performance degradation when evaluating on the target (real) domain. Hence, transforming the depth data from the source (synthetic) into the target domain would be beneficial to the later pseudo ground truth generation. This is an unsupervised adaptation problem where only unaligned data from the source and the target domain is available, as only Y_{Syn} can be accessed while Y_{Real} cannot. We follow similar adversarial adaptation approaches and learn generators as mappings across these domains (see stage 1 in fig. 2). In such a setting, discriminators are employed to enforce similarity between the domain mapping and the respective target domain. This alleviates the need for alignment between both domains. In order to construct the sensor noise model (i.e. synthetic to real domain) correctly, we introduce a min-max normalization η for depth images:

$$\eta(I) = 2 \times \left(\frac{I - \min(I)}{\max(I) - \min(I)} - \frac{1}{2} \right). \qquad (1)$$

By normalizing depth values to lie in the interval $[-1, 1]$ rather than learning in the absolute scale directly, we avoid scale shifting caused by distribution differences among datasets. Additionally, this approach prevents the depth amplitude distribution from becoming the main judging criterion for the discriminator, thereby in turn learning a better sensor noise model. We introduce the sensor noise model N which maps data from the synthetic to the real domain for the purpose of adding realistic noise to clean synthetic samples. It will be optimized to prevent the discriminator D_N from distinguishing between mapped and real depth data. The discriminator, on the other hand, tries to differentiate real noisy data from the mapped ones. We express this objective as:

$$\begin{aligned} &L_{\text{Noise}}(N, D_N, X_{\text{Syn,D}}, X_{\text{Real,D}}) \\ &= \mathbb{E}_{x_t \sim X_{\text{Real,D}}}[log\ D_N(\eta(x_t))] \\ &\quad + \mathbb{E}_{x_s \sim X_{\text{Syn,D}}}[log\ (1 - D_N(\eta(N(\eta(x_s)))))], \end{aligned} \tag{2}$$

where Eq. 2 ensures that N produces convincing sensor-like noisy samples given synthetic clean samples $X_{\text{Syn},D}$. Nonetheless, existing works indicate that networks optimizing such objectives are often unstable, mainly because L_{Noise} does not consider preservation of the original content. Hence a cycle-consistency constraint [48] is imposed on our adaptation procedure. For that purpose, the restoration model R is introduced to map the sensor-like depth map back to the clean synthetic domain, optimising a similar min-max adversarial loss:

$$\begin{aligned} &L_{\text{Restore}}(R, D_R, X_{\text{Real,D}}, X_{\text{Syn,D}}) \\ &= \mathbb{E}_{x_t \sim X_{\text{Syn,D}}}[log\ D_R(\eta(x_t))] \\ &\quad + \mathbb{E}_{x_s \sim X_{\text{Real,D}}}[log\ (1 - D_R(\eta(N(\eta(x_s)))))]. \end{aligned} \tag{3}$$

In contrast to the noise simulator N, the restorer R performs tasks such as hole filling and denoising. More details on how this is accomplished along with qualitative results can be found in the supplementary material. Moreover, an L1 penalty is imposed on samples mapped twice so as to reach their original domain again, e.g. mapping a synthetic sample to the sensor-like depth domain and back to the synthetic domain. This is referred to as the min-max cycle-consistency loss:

$$\begin{aligned} L_{\text{Cycle}}(N, R) = &\mathbb{E}_{x_s \sim X_{\text{Syn,D}}}[\|R(\eta(N(\eta(x_s)))) - \eta(x_s)\|] \\ &+ \mathbb{E}_{x_t \sim X_{\text{Real,D}}}[\|N(\eta(R(\eta(x_t)))) \\ &\qquad - \eta(x_t)\|]. \end{aligned} \tag{4}$$

These three loss functions form our complete objective:

$$\begin{aligned} L(N, R, D_N, D_R) = &L_{\text{Noise}}(N, D_N, X_{\text{Syn,D}}, X_{\text{Real,D}}) \\ &+ L_{\text{Restore}}(R, D_R, X_{\text{Real,D}}, X_{\text{Syn,D}}) \\ &+ L_{\text{Cycle}}(N, R). \end{aligned} \tag{5}$$

Finally, we train the two autoencoders N, R and their respective discriminators, D_N and D_R, jointly by solving the following optimization problem:

$$N, R = \arg\min_{N,R} \max_{D_N, D_R} L(N, R, D_N, D_R). \tag{6}$$

3.2. Training in Adapted Domain

The ability to simulate noise on synthetic training samples enables us in stage two to train a scene parsing model SP_{ada} using the noisy synthetic training samples that mimic the real training samples, denoted $X_{\text{Syn}\to\text{Real,D}} = \{N(\eta(x_{\text{Syn,D}}))\forall x_{\text{Syn,D}} \in X_{\text{Syn,D}}\}$, and the corresponding labels Y_{Syn}. We train the model by minimizing a pixel-wise multinomial logistic regression loss. Additionally, to prevent overfitting towards an unbalanced class distribution, we apply the class balancing strategy proposed in [28]. Formally, the weighted negative log likelihood loss between the prediction and synthetic ground truths for pixel i from a sample $x_{\text{Syn}\to\text{Real,D}}$ can be written as

$$L_{\text{Adapt},i} = -\sum_{c \in C} w_c y_{i,c} \log\left(\frac{e^{p_{i,c}}}{\sum_{c' \in C} e^{p_{i,c'}}}\right), \tag{7}$$

where $p_{i,c}$ is the prediction made by SP_{ada}, $y_{i,c}$ the ground truth label, C the set of classes with weights w_c.

3.3. Pseudo Ground Truth Generation

The third stage utilizes the predictions made by the teacher model SP_{ada} on real depth data and proceeds to generate pseudo ground truth labels Y_{Pseudo}.

3.3.1 Weak Localization

Experiments on depth-only input reveal that the performance of SP_{ada} is still insufficient for certain categories, e.g. books, as their geometry is not distinctive enough. In an attempt to remedy this by adapting a model in the RGB domain, we observed a performance drop nonetheless. This may be due to the domain shift between synthetic and real textures, as a result from the difficulties to model and render certain textures accurately in an automated fashion. Consequently, we propose to utilize weak supervision based on real RGB data as a separate cue for fine-tuning to the object appearance in the target domain. To avoid high labeling costs, we only use image-level tags extracted from SUN RGB-D, without location or boundary information. We generate localization cues by leveraging a CNN that is trained for image classification with a global average pooling layer (GAP) as proposed by [47]. However, the resulting class activity maps (CAM) $\hat{Y}_{\text{CAM}}$ are imprecise, [19] even noted that networks trained with a final GAP overestimate the response region. Hence, we add a 2×2 max pooling layer before the GAP to extract key information and prevent the GAP from overestimation.

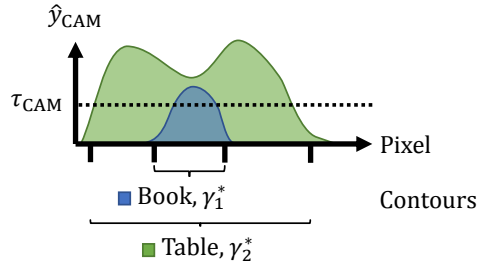

Figure 3. Typical response of the localization heat map for small objects placed on or in larger ones. The response area's size (exceeding τ_{CAM}) is hence used as a decision criterion if there are several confident predictions for a single contour.

3.3.2 Cue Integration

To integrate the depth-based predictions $\hat{Y}_{\text{Adapted}} = SP_{\text{ada}}(X_{\text{Real,D}})$ and the weak localizations $\hat{Y}_{\text{CAM}}$, we propose a two-stage integration mechanism. Our objective is to choose between those two predictions and generate pseudo labels $\hat{Y}_{\text{Pseudo}}$ where we trade coverage for confidence: We prefer learning from fewer but more confident pseudo labels. This trade-off is category-related, different categories have different coverage-confidence profiles that need to be accommodated. We utilize Ultrametric Contour Maps (UCM) [13], a hierarchical representation of the image boundaries, to infer pseudo labels over segments $\gamma_k \in \Gamma$ of the image. We only take those contours into account that exceed a confidence threshold τ_{UCM}, denoting them $\gamma_k^* \in \Gamma^*$.

First Integration Step The first step adds information about the observed geometry to the contours $\gamma_k^* \in \Gamma^*$ by analyzing the depth-based predictions $\hat{Y}_{\text{Adapted}}$ within each contour. In order to remove low confidence labels from $\hat{Y}_{\text{Adapted}}$, we first apply a Softmax and threshold the result against $\tau_{\text{Adapted}} = 0.6$, resulting in $\hat{Y}^*_{\text{Adapted}}$. τ_{Adapted} was chosen so as to balance accuracy and coverage. We then turn to the histogram $H(\gamma_k^*, \hat{y}^*_{\text{Adapted}}) = \{h_{c,k}\}_{c \in \text{Categories}}$ of predicted categories within each contour γ_k^*. Taking a simple maximum likelihood approach, we select the category with the largest histogram value to be the prediction of the first integration step, i.e. $\hat{y}_{\text{Step 1},k} = \arg\max_c h_{c,k}$ for each contour.

Second Integration Step The second integration step decides whether the localization heat maps $\hat{y}_{\text{CAM}}$ provide a more confident prediction than the contours' $\hat{y}_{\text{Step 1},k}$. From $\hat{y}_{\text{CAM}}$ we first generate a proposal set of possible classes P_k for each contour, comprising the most activated class in the heat map and a set of small objects. Since weak localization is provided by a deep neural network that has a rather larger receptive field, small objects may not be accurately represented in the activations. Therefore, we manually add classes that cover only few pixels of an image to the list of candidates to be checked for confidence. Next, we use peak activation $p_{k,c}$ and response rate $r_{k,c}$ to determine which of the proposals is confident enough to replace the estimate from step one, forming the electable set E_k. If this set is empty, i.e. there are no confident localizations, we resort to the result from depth based prediction. Otherwise, if we have several confident classes, we are interested in the category that is most specific, where we take the one with the smallest response area A_c. This way, we avoid neglecting small objects that overlap with larger ones (e.g. books on a table as shown in Fig. 3). All thresholds τ are tuned on 30 random samples of our training set to avoid human learning on the dataset. A listing of the algorithm can be found in the supplementary material.

3.4. Training in Target Domain

After computing Y_{Pseudo} as described in the previous stage, the last stage trains the student network SP_{full} using the real RGB images $X_{\text{Real,RGB}}$ and the estimated labels Y_{Pseudo}. Those estimated labels provide information for only a subset of all pixels, i.e those pixels that we are confident about. In [40] the authors note that they achieved a better disparity map for whole image with only a portion of high-confidence predictions. Hence, assuming that the majority of estimates in Y_{Pseudo} are correct, we expect the missing or incorrectly labeled regions to be recovered by the generalization capability of the neural network. Formally, the loss for the student network for a pixel i is given by

$$L_{\text{Parse},i} = -\sum_{c \in C} w_{\text{Pseudo},c} y_{\text{Pseudo},i,c} \log \left(\frac{e^{p_{i,c}}}{\sum_{c' \in C} e^{p_{i,c'}}} \right) \tag{8}$$

where the pixels in Y_{Pseudo} with unknown or ignored labels do not contribute to the loss. $p_{i,c}$ denotes the prediction for class c at pixel i and $y_{\text{Pseudo},i,c}$ is a one-hot vector containing the pseudo labels.

4. Experiments

To demonstrate the efficacy of our method, several ablation studies on depth-aware adaptation and cue integration are presented. We evaluate our approach on the SUN RGB-D dataset [39]. At first, we present experiments to demonstrate the effectiveness of our depth adaptation method. Two synthetic datasets, SceneNet [25] and Pbrs [45], are used during training procedure. Afterwards we proceed to ablation studies of our model to show the effect of each measure on the final result. Finally, we compare to state-of-the-art domain adaptation scene parsing methods and their fully-supervised counterpart. Note that, to be more realistic, no additional real data is used. The only annotations used

Table 1. Ablation study of minmax normalization for depth adaptation. Results reported are from the SUN RGB-D semantic segmentation validation set. Best values are highlighted in bold font.

	bed	books	ceil	chair	floor	furn.	objs.	paint	sofa	table	tv	wall	mIoU (w/o windows)
Ours Depth (Raw, w/o minmax normalization)	27.85	0.00	28.36	26.37	72.29	24.84	10.91	4.13	23.28	34.21	**6.23**	58.78	26.44
Ours Depth (Raw)	**40.20**	0.00	**33.77**	**31.21**	**72.30**	**30.06**	**11.61**	**13.02**	**31.75**	**40.13**	4.49	**62.81**	**30.95**

Table 2. Ablation study of sensor noise simulation. These results were reported on both inpainted and raw SUN RGB-D validation set.

	bed	books	ceil	chair	floor	furn.	objs.	paint	sofa	table	tv	wall	mIoU (w/o windows)
Inpainted													
Syn Depth	33.06	0.00	25.86	24.42	76.22	26.70	9.85	9.74	26.22	38.70	6.36	**63.91**	28.42
Ours Depth (w/o Cycle Loss)	33.32	0.00	32.07	31.76	71.13	25.71	12.73	10.02	32.09	36.50	6.33	53.88	28.80
[5]+Syn Depth	38.55	0.00	**37.60**	41.21	78.25	28.28	12.80	**16.26**	29.41	39.71	5.85	63.34	32.61
Ours Depth	**49.04**	0.00	35.75	**41.40**	**79.55**	**31.44**	**14.68**	14.63	**38.51**	**43.73**	**7.78**	61.83	**34.86**
Raw													
Syn Depth	25.92	0.00	31.37	18.97	54.30	22.25	6.95	8.22	19.40	29.24	2.96	47.02	22.22
[5] Depth	30.31	0.00	33.54	22.89	**72.40**	26.43	11.11	13.01	25.54	36.34	**4.57**	61.12	28.11
Ours Depth(Raw)	**40.20**	0.00	**33.77**	**31.21**	72.30	**30.06**	**11.61**	**13.02**	**31.75**	**40.13**	4.49	**62.81**	**30.95**

in our setting are image-level tags from the SUN RGB-D dataset which are much cheaper to acquire than pixel-wise annotations or object bounding boxes.

4.1. Implementation Details

All experiments are implemented in the Pytorch 0.3 [29] framework with CUDA 9.0 and CuDNN backends on a single NVIDIA Titan X. For a fair comparison and consideration of computational efficiency, we evaluate our approaches and the state-of-the-art adaptation method CYCADA [15] using the ERFNet [34] network architecture. Without loss of generality, our method can be applied to other scene parsing models. Our reproduction of CYCADA is trained with the hyperparameters as published by the authors, including weight sharing. For the scene parsing model, the input images were resized to 320×240 and the Adam [18] variant of stochastic gradient descent is used for minimization of all loss functions. Training is performed with a batch size of 48. Moreover, we train with an initial learning rate of 5×10^{-4} and reduce it by half once loss value stalls so as to accelerate convergence as done in [34]. We apply standard data augmentation techniques like dropout, random flipping and cropping to prevent our models from overfitting. For the weakly supervised model, we use the encoder of ERFNet pretrained on ImageNet [11] for initialization and replace the original fully-connected layers with a max-pooling, a global average pooling and a new fully-connected softmax layer.

4.2. Ablation Studies

Minmax adversarial loss Table 1 demonstrates the effects of minmax normalization on sensor noise learning. The IoU of most categories is improved significantly in the minmax normalization setting over raw depth data. This shows the utility of the normaliser η in suppressing depth magnitude based learning in the discriminators D_N and D_R. Note that the loss of category window is set to zero and excluded in this comparison due to wrong depth reported by the active sensors employed in creating the SUN RGB-D dataset.

Sensor depth simulation To show the efficacy of sensor depth simulation, we evaluate the performance of depth-based scene parsing as shown in Table 2. Our evaluation includes both the raw as well as the inpainted depth maps as provided by the SUN RGB-D. We compare with models trained solely from synthetic data as well as to the simulation method proposed in [5]. Our method clearly outperforms those two methods for both kinds of depth maps, thereby establishing a new baseline for our following adaptation experiments.

Cues and Integration Table 3 disentangles the influence of individual cues and integration mechanisms in each voting stage during training. The results show how both integration stages contribute to the IoU improvement in different categories to various extent, thereby complementing each other. Class heat maps are particularly helpful for those objects that are smaller or do not possess a distinctive geometric structure. Note that while the overall mIoU performance improves, the mIoU for some categories such as bed and table degrades after the second integration step due to the inter-class occurrence issue: In weak localization where no explicit object positions are available, those cate-

Table 3. Influence of cues and voting mechanism in each stage. These results were obtained on SUN RGB-D validation set.

	Input	Training Label	bed	books	ceil	chair	floor	furn.	objs.	paint	sofa	table	tv	wall	window	**mIOU**
w/o Adaptation	RGB	Y_{Syn}	22.57	0.00	46.84	42.50	62.82	24.55	13.86	18.96	31.81	27.45	5.76	55.74	28.58	29.34
Ours Depth	Depth	Y_{Syn}	49.04	0.00	35.75	41.40	79.55	31.44	14.68	14.63	38.51	43.73	7.78	61.83	0.91	32.25
Ours(1st stage only)	RGB	$\hat{Y}_{\text{Step\_1}}$	**54.93**	0.00	**53.12**	47.50	79.64	35.77	15.99	0.00	40.39	**48.89**	16.07	64.82	0.65	35.21
Ours(2nd stage only)	RGB	UCM+$\hat{Y}_{\text{CAM}}$	27.71	12.87	16.13	36.19	29.17	13.12	12.95	20.15	34.56	31.27	7.81	50.72	**44.99**	25.97
Ours(Full)	RGB	Y_{Pseudo}	52.06	**23.52**	50.03	**49.44**	**81.00**	**36.39**	**25.17**	**28.09**	**44.64**	47.88	**19.68**	**69.69**	38.25	**43.53**

Table 4. Comparison of pseudo labels Y_{Pseudo} to our final model. Quantities labeled "effective" refer to the original quantity multiplied by the cover ratio, thereby taking only valid pixels into account for a more accurate comparison. Those labeled @Y_{Pseudo} are evaluated only on those pixels where pseudo labels Y_{Pseudo} are available. Evaluations are conducted on the SUN RGB-D dataset. GA refers to the Global Accuracy over all pixels.

Predictions	Dataset partition	Cover ratio	GA	GA@Y_{Pseudo}	Effective GA	mIoU	mIoU@Y_{Pseudo}	Effective mIoU
Y_{Pseudo}	Training	72.77	80.86	80.86	58.84	56.97	56.97	41.64
SP_{full}	Training	100	75.89	80.91	75.89	49.46	56.74	49.46
SP_{full} (incl. UCM refinement)	Training	97.73	76.81	81.29	75.07	50.81	57.52	49.66
SP_{full}	Validation	100	73.64	-	73.64	43.53	-	43.53

gories that mostly appear together in a scene cannot always be properly separated, i.e. their labels could be swapped without invalidating any data.

Student Network Table 4 compares the result of the student network with the pseudo labels Y_{Pseudo}, i.e. evaluating on Y_{Pseudo} directly without training the student network. This illustrates how the student network is able to learn a scene parsing model that is more accurate than its training data. In order to evaluate performance matrix and cover ratio, i.e., percentage of valid pixels simultaneously, the quantities labelled effective in the table refers to multiplying both. Note that effective mIoU is calculated by using class-wise cover ratio instead of global ones.

Combination of Cues Table 5 shows additional experiments for different cues. We realised perfect depth transfer by letting the supervised model generate the pseudo labels $\hat{y}_{\text{Adapted}}$ from depth input. As expected, our results lie in-between, i.e. our depth transfer enables improvements but cannot fully compete with supervised information. In addition, we swapped the data and adapted RGB while applying weak supervision to the depth cue. The result is almost five percentage points below ours. Adapting RGB and using it in weak supervision at the same time brings the result closer to ours, however mostly due to improvements in the category "window", while our approach performs better in most other categories. This indicates that synthetic RGB data may not be necessary, which can reduce the dataset creation effort as texturing, lighting etc. can be avoided.

4.3. Comparison to the State-of-the-Art

In Table 6, we compare our results to full supervision and the state-of-the-art domain adaptation methods. For a fair comparison, all models, including CYCADA are trained using the ERFNet architecture. The not adapting alternative, denoted NADA, is trained on synthetic data directly. CYCADA, the state-of-the-art domain adaption method, was trained starting from the pretrained NADA parameters. Although CYCADA outperforms NADA and performs better than our depth adaptation on categories with indistinctive geometric structures such as paint, tv, windows, there is only a slight improvement, which comes at the high effort of computer generated imagery. It shows that taking appearance from real data into account yields significant advantages even if only image-level labels are available. Fig. 4 shows examples of our final result. Note that some visualization of our results seems to be incorrect when compared to the ground truth. However, we observed that some ground truths are imprecise and a portion of regions marked "unknown" can be predicted correctly if we align our result with RGB inputs by applying simple UCM based contour-wise voting using predictions at inference phase. Hence, we argue that the performance of our approach may still be underestimated by this evaluation. We provide further examples for this phenomenon in the supplementary material.

5. Conclusions

Starting out from synthetically generated scene parsing data, we have demonstrated how transferring information in the depth domain can exploit the smaller domain gap of geometric data for indoor scene parsing. Proceeding to integrate weak localization can recover information that is not present or difficult to detect in synthetic indoor scenery. Altogether this yields a significant performance improvement for learning indoor scene parsing without dense labels, reducing the mIoU drop from 47% to 20%. While we utilize depth for our adaptation, this is only necessary at training, not at inference time, thereby maintaining a low computations and sensory footprint. These techniques may readily applied and extended to benefit other computer vision tasks in the future.

Table 5. Influence of RGB and depth cues. These results were obtained on SUN RGB-D validation set.

Input	bed	books	ceil	chair	floor	furn.	objs.	paint	sofa	table	tv	wall	window	**mIOU**
Ours, D (adaptation) + RGB (weak)	52.26	23.52	50.03	49.44	81.00	36.39	25.17	28.29	44.64	47.88	19.68	69.69	38.25	**43.53**
Depth (w/o adaptation) + RGB (weak)	45.52	15.32	40.35	44.44	77.87	38.00	23.12	26.83	44.54	46.24	16.24	68.79	38.94	**40.48**
Depth (perfect transfer) + RGB (weak)	54.48	20.62	57.69	52.75	83.27	43.61	33.15	32.30	48.46	53.11	16.07	73.61	50.94	**47.77**
RGB (adaptation) + D (weak)	51.77	16.10	47.42	47.54	77.31	28.24	15.87	22.89	44.59	46.72	0.00	62.08	43.64	**38.78**
RGB (adaptation) + RGB (weak)	48.53	19.64	48.14	48.27	77.58	36.34	23.22	29.46	44.59	47.45	21.66	68.68	47.68	**43.17**

Table 6. Comparison of our approach to state-of-the-art domain adaptation and fully-supervised methods. Results are obtained on the SUN RGB-D validation set.

Method	Dataset SUN	Dataset Scene	Pbrs	bed	books	ceil	chair	floor	furn.	objs.	paint	sofa	table	tv	wall	window	mIOU	mIOU drop (rel.)
Supervised [34]	√ (full)			62.46	26.07	67.54	62.52	85.68	47.10	38.43	43.15	49.72	59.33	40.49	76.92	54.12	54.89	-
NADA [45]			√	22.13	0.00	23.42	40.08	69.58	23.70	10.34	5.05	36.38	21.90	8.97	57.15	23.27	26.31	-52.07%
CYCADA [15]			√	28.22	0.00	24.39	39.57	68.45	23.51	12.61	15.42	39.00	16.65	13.74	59.12	34.95	28.90	-47.35%
Ours Depth			√	48.11	0.00	22.24	39.99	77.18	27.59	13.92	12.01	39.35	39.32	6.34	59.08	0.00	29.24	-46.73%
Ours Depth		√	√	49.04	0.00	35.75	41.40	79.55	31.44	14.68	14.63	38.51	43.73	7.78	61.83	0.91	32.25	-41.25%
Ours (Full)	√ (weak)	√	√	52.06	**23.52**	**50.03**	49.44	81.00	36.39	**25.17**	28.09	44.64	47.88	**19.68**	69.69	38.25	43.53	-20.70%
Ours (Full +UCM Refinement)	√ (weak)	√	√	**54.07**	21.94	47.54	**50.37**	**81.10**	**36.56**	24.75	**30.67**	**46.23**	**49.15**	17.76	**70.19**	**39.00**	**43.80**	**-20.20%**

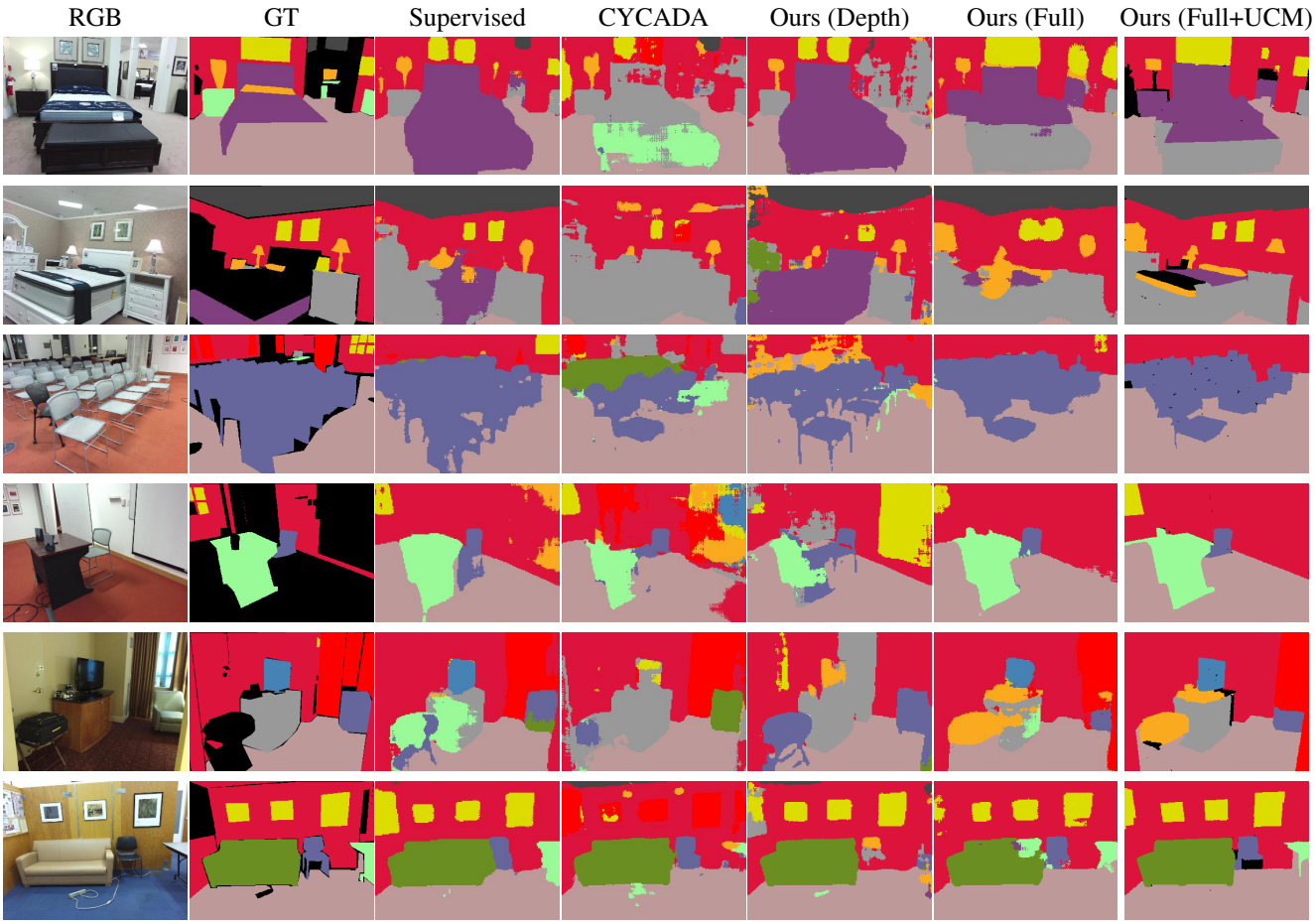

Figure 4. Visualization and comparison of our method. Note that UCM helps aligning the predictions with image boundaries. Overlaid images for more examples are shown in the supplementary.

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. *CoRR*, abs/1803.10464, 2018.

[2] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems 15 Neural Information Processing Systems, NIPS*, pages 561–568, 2002.

[3] Amy L. Bearman, Olga Russakovsky, Vittorio Ferrari, and Fei-Fei Li. What's the point: Semantic segmentation with point supervision. In *Computer Vision - ECCV 2016 - 14th European Conference*, pages 549–565, 2016.

[4] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, pages 41–48, 2009.

[5] Jeannette Bohg, Javier Romero, Alexander Herzog, and Stefan Schaal. Robot arm pose estimation through pixel-wise part classification. In *2014 IEEE International Conference on Robotics and Automation, ICRA*, pages 3143–3150, 2014.

[6] Arslan Chaudhry, Puneet Kumar Dokania, and Philip H. S. Torr. Discovering class-specific pixels for weakly-supervised semantic segmentation. In *British Machine Vision Conference 2017, BMVC*, 2017.

[7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017.

[8] Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *IEEE International Conference on Computer Vision, ICCV*, pages 2011–2020, 2017.

[9] Yuhua Chen, Wen Li, and Luc Van Gool. ROAD: reality oriented adaptation for semantic segmentation of urban scenes. *CoRR*, abs/1711.11556, 2017.

[10] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *2015 IEEE International Conference on Computer Vision, ICCV*, pages 1635–1643, 2015.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR*, pages 248–255, 2009.

[12] Mark Everingham, S. M. Ali Eslami, Luc J. Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.

[13] Saurabh Gupta, Ross B. Girshick, Pablo Andrés Arbeláez, and Jitendra Malik. Learning rich features from RGB-D images for object detection and segmentation. In *Computer Vision - ECCV 2014 - 13th European Conference*, pages 345–360, 2014.

[14] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

[15] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *Proceedings of the 35th International Conference on Machine Learning, ICML*, pages 1994–2003, 2018.

[16] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *CoRR*, abs/1612.02649, 2016.

[17] H. J. Scudder III. Probability of error of some adaptive pattern-recognition machines. *IEEE Trans. Information Theory*, 11(3):363–371, 1965.

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[19] Alexander Kolesnikov and Christoph H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *Computer Vision - ECCV 2016 - 14th European Conference*, pages 695–711, 2016.

[20] Suha Kwak, Seunghoon Hong, and Bohyung Han. Weakly supervised semantic segmentation using superpixel pooling network. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 4111–4117, 2017.

[21] Qizhu Li, Anurag Arnab, and Philip H. S. Torr. Weakly- and semi-supervised panoptic segmentation. In *Computer Vision - ECCV 2018 - 15th European Conference*, pages 106–124, 2018.

[22] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 3159–3167, 2016.

[23] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian D. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 5168–5177, 2017.

[24] Lingni Ma, Jörg Stückler, Christian Kerl, and Daniel Cremers. Multi-view deep learning for consistent semantic mapping with RGB-D cameras. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, pages 598–605, 2017.

[25] John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J. Davison. Scenenet RGB-D: 5m photorealistic images of synthetic indoor trajectories with ground truth. *CoRR*, abs/1612.05079, 2016.

[26] Seong Joon Oh, Rodrigo Benenson, Anna Khoreva, Zeynep Akata, Mario Fritz, and Bernt Schiele. Exploiting saliency for object segmentation from image level labels. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 5038–5047, 2017.

[27] George Papandreou, Liang-Chieh Chen, Kevin P. Murphy, and Alan L. Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *2015 IEEE International Conference on Computer Vision, ICCV*, pages 1742–1750, 2015.

[28] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *CoRR*, abs/1606.02147, 2016.

[29] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

[30] Deepak Pathak, Philipp Krähenbühl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *2015 IEEE International Conference on Computer Vision, ICCV*, pages 1796–1804, 2015.

[31] Deepak Pathak, Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional multi-class multiple instance learning. *CoRR*, abs/1412.7144, 2014.

[32] Pedro H. O. Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1713–1721, 2015.

[33] Xiaojuan Qi, Zhengzhe Liu, Jianping Shi, Hengshuang Zhao, and Jiaya Jia. Augmented feedback in semantic segmentation under image level supervision. In *Computer Vision - ECCV 2016 - 14th European Conference*, pages 90–105, 2016.

[34] Eduardo Romera, Jose M. Alvarez, Luis Miguel Bergasa, and Roberto Arroyo. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Trans. Intelligent Transportation Systems*, 19(1):263–272, 2018.

[35] Anirban Roy and Sinisa Todorovic. Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 7282–7291, 2017.

[36] Fatemehsadat Saleh, Mohammad Sadegh Ali Akbarian, Mathieu Salzmann, Lars Petersson, and Jose M. Alvarez. Bringing background into the foreground: Making all classes equal in weakly-supervised video semantic segmentation. In *IEEE International Conference on Computer Vision, ICCV*, pages 2125–2135, 2017.

[37] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision, ICCV*, pages 618–626, 2017.

[38] Wataru Shimoda and Keiji Yanai. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In *Computer Vision - ECCV 2016 - 14th European Conference*.

[39] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 567–576, 2015.

[40] Alessio Tonioni, Matteo Poggi, Stefano Mattoccia, and Luigi di Stefano. Unsupervised adaptation for deep stereo. In *IEEE International Conference on Computer Vision, ICCV*, pages 1614–1622, 2017.

[41] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. *CoRR*, abs/1806.04659, 2018.

[42] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. *CoRR*, abs/1703.08448, 2017.

[43] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S. Huang. Revisiting dilated convolution: A simple approach for weakly- and semi- supervised semantic segmentation. *CoRR*, abs/1805.04574, 2018.

[44] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *IEEE International Conference on Computer Vision, ICCV*, pages 2039–2049, 2017.

[45] Yinda Zhang, Shuran Song, Ersin Yumer, Manolis Savva, Joon-Young Lee, Hailin Jin, and Thomas A. Funkhouser. Physically-based rendering for indoor scene understanding using convolutional neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5057–5065, 2017.

[46] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 6230–6239, 2017.

[47] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 2921–2929, 2016.

[48] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision, ICCV*, pages 2242–2251, 2017.